Z-113E
C.2.

Statistics Canada
Statistique Canada

Ottawa

NONPARAMETRIC BAYESIAN MODELS FOR SAMPLES FROM FINITE POPULATIONS

by

DAVID A. BINDER[1]

December 1981

[1]*Some research for this paper was conducted while the author was in the Faculty of Administration, University of Ottawa*

Summary

Using Ferguson's (1973) non-parametric priors in a Bayesian analysis of finite populations, we show that asymptotically, at least, the usual estimates and confidence intervals for the population mean in simple and stratified random samples can be justified in Bayesian terms. We then apply these models for estimating population percentiles, and a new procedure for interval estimates in stratified sampling is developed.

Keywords: Asymptotic posteriors; Dirichlet process; Population mean; Population percentiles; Stratified samples; Superpopulation models.

1. INTRODUCTION

At present there are a number of general approaches in the statistical literature for making inferences based on data selected from a finite population. Many of these are summarized in Smith (1976). Probably the most widely used methodology, in practice, is the design-based inference from random samples, where criteria such as bias and mean-squared error, derived from the sampling distribution of the data, are used. This methodology does not preclude the possibility that the estimator may be used on an inherent superpopulation model; however, confidence intervals and mean-squared errors are based only on the probability distribution of the design, and ignore the model. One of the reasons for the popularity of this approach may be that it is nonparametric in nature and sample sizes tend to be large for many applications, so that the loss of efficiency compared with parametric superpopulation approaches is considered less important than the robustness of the design-based approach to a wide class of models.

A Bayesian approach to inference in a finite population with exchangeable priors was introduced by Ericson (1969). In order to facilitate the descriptions of the exchangeable prior, Ericson (1969) suggests using parametric superpopulation models; that is, given a vector parameter θ, the finite population can be assumed to be the realization of independent and identically distributed observations

each with distribution function $P(\cdot|\theta)$. For any prior, on θ, this yields an exchangeable prior on the finite population values. Ericson (1969) points out that a disadvantage of this approach is that strong prior knowledge regarding the shape of the finite population distribution is assumed. To avoid this he suggests assuming the superpopulation distribution is discrete and he uses a Dirichlet prior distribution as a convenient approximation. Scott (1971) has shown that under suitable regularity conditions, the asymptotic posterior for the values of the finite population is independent of the prior and is related to the multivariate normal distribution.

One drawback of this approach is that the discrete values of the superpopulation distribution are assumed to be a subset of some known countable set. In this paper we extend this to allow the discrete values to take any real value by using nonparametric priors, introduced by Ferguson (1973); namely the Dirichlet process. We examine the estimation of population means and percentiles for simple and stratified random sampling.

In Section 2 we briefly review the Dirichlet process and quote the results for estimating population means.

Section 3 discusses the implication of these models to estimating population percentiles. Analogies with the design-based approach in Woodruff (1952) and Sedransk and Meyer (1978) are given.

We will not address the problem of how samples should be selected in this paper. Since our approach is Bayesian, we do not need to know the probability distribution of the design. We concentrate only on the inferences we can make on the finite population for a given sample.

2. THE MODEL AND RESULTS FOR POPULATION MEANS

2.1 Dirichlet Process

One of the simplest outlines of the Dirichlet process, is given by Breth (1979):

"Let $\mathcal{X}$ be the space of the possible values, and let $(A_1, \ldots, A_{m+1})$ be a collection of events, one and only of which must occur, i.e. a partition. Write $P(A_i)$ for the random probability, $pr(A_i)$. We have a Dirichlet process if for every possible partition for all possible m, the random multinomial probabilities $\{P(A_1), \ldots, P(A_m)\}$ have a Dirichlet distribution with parameter $\{\alpha(A_1), \ldots, \alpha(A_m); \alpha(A_{m+1})\}$ (Wilks, 1962, p. 178)".

To denote this process, we use the notation $P \varepsilon \mathcal{D}(\alpha)$ for a random probability distribution arising from a Dirichlet process. This process was studied by Ferguson (1973, 1974). Now, if $P \varepsilon \mathcal{D}(\alpha)$ and we observe $X_1, \ldots, X_n$, independent and identically distributed random variables with distribution P, then the posterior for P is a Dirichlet process with parameter $\alpha + \Sigma \delta_{x_i}$, where δ_x denotes the measure giving mass one to the point x.

A disadvantage of the Dirichlet process prior is that if $P \varepsilon \mathcal{D}(\alpha)$, then, with probability one, P is a discrete distribution. However, for finite populations this is not a serious restriction since as Ericson (1969, p. 206) says, there is an "inherent discreteness of almost all observations due to limitations of measuring instruments, etc." Scott (1971) regards this discretenss as corresponding to the "realities of practice".

2.2 Single Population

We assume there is a single superpopulation distribution from which all the population values are derived. Suppose $Y_1, \ldots, Y_N$ are the population values for a finite population of size N, and given a superpopulation distribution P, these values are independent and identically distributed, with distribution P.

Now, $y_1, \ldots, y_n$ are the sample values from this population, and we study the case where the prior for P is a simple Dirichlet process with parameter α. We wish to make inferences on $\mu = \frac{1}{N} \Sigma Y_i$, given the observations $y_1, \ldots, y_n$. We let $\bar{y} = n^{-1} \Sigma y_i$ and $s^2 = n^{-1} \Sigma (y_i - \bar{y})^2$.

Before taking any observations, the prior mean for μ is denoted by $\hat{\mu}_o$ and the prior variance for μ is $\sigma_o^2(\mu)$. We let $A = \alpha(-\infty,\infty)$, the value of the Dirichlet process parameter on the whole real line. We find that the posterior mean for μ is

$$\hat{\mu}_n = N^{-1}(A+n)^{-1}\{(N+A)\ n\ \bar{y} + (N-n)\ A\ \hat{\mu}_o\}. \tag{2.1}$$

The posterior variance for μ is

$$\sigma_n^2(\mu) = \frac{(N-n)(A+N)}{N^2(A+n)(A+n+1)}\left\{ns^2 + \frac{nA}{(n+A)}\ (\bar{y}-\hat{\mu}_o)^2\right\} + \frac{(N-n)\ A(A+1)}{N(A+n)(A+n+1)}\ \sigma_o^2(\mu) \tag{2.2}$$

We see that for large sample sizes

$$\hat{\mu}_n \doteq \bar{y} \tag{2.3}$$

$$\sigma_n^2(\mu) \doteq \frac{N-n}{N(n+1)}\ s^2. \tag{2.4}$$

Expressions (2.1) to (2.4) appear in Ericson (1969; pp. 212, 213), where it is assumed that the superpopulation is discrete with probability masses at known values.

We consider now the asymptotic posterior distribution for μ. We suppose that there are k distinct values among $y_1, \ldots, y_n$ and we form the partition $(A_1, \ldots, A_k)$ such that each distinct value is a member of exactly one of the A_i's. We let z_i be the sampled value in A_i, and $n(A_i) = np_i$ be the number of sampled values equal to z_i. If Y_j^* is an unsampled value of the population, we obtain that in the posterior distribution $pr(Y_j^* = z_i | Y_j^* \in A_i) = n(A_i)/\{\alpha(A_i) + n(A_i)\}$, so that if the superpopulation distribution is discrete, this tends to one for large n. For $N(A_i)$ defined as the number of population values in A_i, we thus have that asymptotically

$$\mu = N^{-1}\ \Sigma\ N(A_i)\ z_i$$

$$\hat{\mu}_n = \Sigma p_i\ z_i.$$

Scott (1971) has shown that as $n \to \infty$ and $N-n \to \infty$ and for $U_i = n^{\frac{1}{2}} N^{-\frac{1}{2}}(N-n)^{-\frac{1}{2}}$ $[N(A_i) - Np_i]$, we have that $(U_1, \ldots, U_k)$ is asymptotically singular multivariate normal with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$ where

$$\sigma_{ij} = p_i(1-p_i) \text{ if } i=j$$
$$= -p_i p_j \quad \text{if } i \neq j.$$

Therefore, we have $\left\{\frac{nN}{N-n}\right\}^{\frac{1}{2}} (\mu - \hat{\mu}_n)$ is $N(0,s^2)$ as $n \to \infty$ and $N-n \to \infty$. We see that Bayesian prediction intervals are asymptotically equivalent to the usual confidence intervals for this case.

2.3 Stratified Population

We now assume that the population values are stratified into L strata, with sizes $N_1, \ldots, N_L$. The strata may have been incorporated into the original design, or the observations may have been post-stratified. The population values for the i-th stratum are $Y_{i1}, \ldots, Y_{iN_i}$. The sampled values are $y_{i1}, \ldots, y_{in_i}$.

We assume that in the i-th stratum, the population values are independent and identically distributed with probability function P_i. We study the case where the prior for $(P_1, \ldots, P_L)$ is such that the probability functions are independent and $P_i \in \mathcal{D}(\alpha_i)$.

Now the posterior for $(P_1, \ldots, P_L)$ given $\{y_{ij}\}$ is that the probability functions are independent and $P_i \in \mathcal{D}(\alpha_i + \Sigma_j \delta_{y_{ij}})$.

We define $\mu_i = N_i^{-1} \Sigma_j Y_{ij}$ and we wish to make inferences on $\mu = N^{-1} \Sigma N_i \mu_i$. Analogously to Section 2.2, we let $A_i = \alpha_i(-\infty,\infty)$, $\bar{y}_i = n_i^{-1} \Sigma_j y_{ij}$, $s_i^2 = n_i^{-1} \Sigma_j (y_{ij} - \bar{y}_i)^2$. The prior expectation of μ_i is denoted by $\hat{\mu}_{io}$, and the prior variance of μ_i is $\sigma_o^2(\mu_i)$. The posterior expectation and variance of μ is

$$\hat{\mu}_n = \Sigma_i \left\{\frac{(N_i+A_i)\, n_i \bar{y}_i + (N_i-n_i)\, A_i \hat{\mu}_{io}}{N(n_i+A_i)}\right\}, \text{ and}$$

$$\sigma^2(\mu) = \Sigma_i \left[\frac{(N_i-n_i)(N_i+A_i)}{N^2(n_i+A_i)(n_i+A_i+1)} \left\{n_i s_i^2 + \frac{n_i A_i}{(n_i+A_i)} (\bar{y}_i - \hat{\mu}_{io})^2\right\} + \frac{(N_i-n_i)\, A_i(A_i+1)}{N_i(A_i+n_i)(A_i+n_i+1)} \sigma_o^2(\mu_i)\right.$$

For large sample sizes, we have

$$\hat{\mu}_n \doteq N^{-1} \Sigma N_i \bar{y}_i$$

$$\sigma^2(\mu) \doteq N^{-2} \Sigma\{(N_i-n_i)\, N_i s_i^2/(n_i+1)\},$$

in the posterior.

Also, if $n_i \to \infty$, $N_i - n_i \to \infty$, we observe that

$$\left\{\Sigma_i \frac{(N_i-n_i)\, N_i s_i^2}{N^2\, n_i}\right\}^{\frac{1}{2}} \{\mu - \hat{\mu}_n\}$$

is asymptotically N(0,1). Again, for large samples, the Bayesian prediction intervals correspond to the usual confidence intervals from stratified samples.

3. POPULATION PERCENTILES

3.1 Single Population

In this section we study the problem of estimating the population percentiles.

We first consider the case where $P \in \mathcal{D}(\alpha)$ and given P, the values $Y_1, \ldots, Y_N$ are independent and identically distributed with distribution P. We order the population values and re-label them $Y_{(1)}, \ldots, Y_{(N)}$, so that $Y_{(1)} \leq \ldots \leq Y_{(N)}$. We study the problem of estimating $Y_{(i)}$ based on the sample $y_{(1)}, \ldots, y_{(n)}$, where $y_{(1)} \leq \ldots \leq y_{(n)}$. Campbell and Hollander (1978) consider a similar problem with Dirichlet process priors. They were interested in estimating the rank of $y_{(j)}$ among $Y_{(1)}, \ldots, Y_{(N)}$. Hill (1968) considers this problem under a different prior distribution. Ericson (1969) uses the Dirichlet prior where the discrete values of the population are known. Our results are analogous to Ericson's (1969) for the exact posterior distribution. We extend these to the asymptotic case. Sedransk and Meyer (1978) consider this problem in a non-Bayesian context, developing confidence intervals of the form $y_{(j)} \leq Y_{(i)} \leq y_{(k)}$. Woodruff (1952) developed large-sample confidence intervals for $Y_{(i)}$.

Now, the prior distribution for $Y_{(i)}$ in our Dirichlet process model is given by:

$$\text{pr}\{Y_{(i)} \leq y\} = \text{pr}\{U(y) \geq i\},$$

where U(y) is the number of population values less than or equal to y. If we define $A(y) = \alpha(-\infty, y]$ and $A = \alpha(-\infty, \infty)$, then U(y) has a Beta-Binomial $[N; A(y), A-A(y)]$ distribution.

The Beta-Binomial $(M; \beta_1, \beta_2)$ distribution is given by

$$f(x) = \binom{M}{x} \frac{\Gamma(\beta_1+\beta_2)\ \Gamma(\beta_1+x)\ \Gamma(\beta_2+M-x)}{\Gamma(M+\beta_1+\beta_2)\ \Gamma(\beta_1)\ \Gamma(\beta_2)}; \quad x=0,1, \ldots, M.$$

Suppose we observe $y_{(1)}, \ldots, y_{(n)}$. By convention we let $y_{(o)} = -\infty$. We take $Y^*_{(1)}, \ldots, Y^*_{(N-n)}$ to be the order statistics for the unsampled units. For $j<i$ and $y_{(j)} \leq y < y_{(j+1)}$,

$$\begin{aligned} \text{pr}[Y_{(i)} \leq y] &= \text{pr}[Y^*_{(i-j)} \leq y] \\ &= \text{pr}(U(y) \geq i-j), \end{aligned} \tag{3.1}$$

where $U(y) \sim$ Beta-Binomial $[N-n;\ A(y) + j,\ A-A(y) + n-j]$. This gives the exact posterior distribution for $Y_{(i)}$. Ericson (1969) has demonstrated that with diffuse priors; that is, $A(y) \doteq 0$, the Bayesian prediction intervals correspond closely with the classical confidence intervals, providing N is large relative to n.

Now if $n \to \infty$, $N-n \to \infty$, $j/n \to p$, then by the asymptotic normality of the Beta-Binomial distribution derived as a special case of the theorem in Scott (1971), we have that expression (3.1) is asymptotically

$$\text{pr}\left[W \geq \left\{\frac{nN}{(N-n)\ p(1-p)}\right\}^{\frac{1}{2}} \left(\frac{i}{N} - p\right)\right], \text{ where W is standard normal.}$$

The resulting prediction intervals are analogous to Woodruff's (1952) large-sample confidence intervals. The two intervals would be identical if p(1-p) in the first term were replaced by $i(N-i)/N^2$. For large samples the difference between the two intervals would be negligible.

3.2 Stratified Populations

Sedransk and Meyer (1978) have considered the problem of developing confidence intervals for population percentiles in stratified random sampling. They consider the case L = 2 in detail and suggest a number of alternative procedures. They resort to a Monte Carlo study to assess which procedure appears to perform well. Their procedure becomes very complex when more than two strata are involved.

Proceeding as in Section 2.3, we assume the $P_i \in \mathcal{D}(\alpha_i)$ and that $(P_1, \ldots, P_2)$ are independent. We obtain $\{Y_{ij};\ 1 \le i \le L,\ 1 \le j \le N_i\}$ which, given $(P_1, \ldots, P_L)$, are independent with Y_{ij} having distribution P_i. From the i-th stratum, we have observations $y_{i1}, \ldots, y_{in_i}$.

We define $Z_{(1)}, Z_{(2)}, \ldots, Z_{(N_1 + \ldots + N_L)}$ to be the order statistics for all the population values, without regard to stratum, where $Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(N_1 + \ldots + N_L)}$.
We define $y_{i,(1)}, \ldots, y_{i,(n_i)}$ to be the order statistics from the i-th stratum. We let $Z^*_{(1)}, \ldots, Z^*_{(\Sigma N_i - \Sigma n_i)}$ be the order statistics of the unsampled population values. We are interested in the posterior probability that $Z_{(j)} \le z$ given all the y's. We define $U_i(z)$ to be the number of population values in stratum i which are less than or equal to z. Therefore.

$$\text{pr}\{Z_{(j)} \le z\} = \text{pr}\{\Sigma U_i(z) \ge j\}.$$

For a given value of z, we define $p_1, \ldots, p_L$ to satisfy

$$y_{i,(n_i p_i)} \le z < y_{i,(n_i p_i + 1)};$$

that is, p_i is the proportion of observations in the i-th stratum which are less than or equal to z. Now, in the posterior distribution, for $j > \Sigma\, n_i p_i$

$$\text{pr}\{Z_{(j)} \le z\} = \text{pr}\{Z^*_{(j - \Sigma n_i p_i)} \le z\}$$

$$= \text{pr}\{\Sigma\, U_i(z) \ge j - n_i p_i\}, \qquad (3.2)$$

where $U_1, \ldots, U_L$ are independent with

$$U_i \sim \text{Beta-Binomial}\{N_i - n_i;\ A_i(z) + n_i p_i,\ A_i - A_i(z) + n_i(1-p_i)\}.$$

We now compare the small-sample prediction intervals with the confidence intervals for population percentiles developed by Sedransk and Meyer (1978). In particular, we consider $L = 2$ strata and for simplicity we assume, analogous to Ericson (1969), that all sample values are distinct and $A(y) \doteq 0$. We define $Y_{i,(j)}$ as the j-th order statistic from the i-th stratum on the quantity $\text{pr}\{Z_{(t)} \le z \mid Z_{(t)} = Y_{i,(j)}\}$. Once these are derived, the confidence intervals depend on various assumptions on the stratum values in order to compute $\text{pr}\{Z_{(t)} = Y_{i,(j)}\}$.

We compute the Bayes' posterior probability for the event $Y_{i,(j)} \le z$, where $z = y_{i,(m_i)}$. Proceeding as above, we see that this probability is the same as $\text{pr}\{U_i(z) \ge j - m_i\}$, where U_i is Beta-Binomial $\{N_i - n_i;\ m_i,\ n_i - m_i\}$. Therefore,

$$\text{pr}\{Y_{i,(j)} \le z\} = \sum_{u=j-m_i}^{N_i-n_i} \binom{u+m_i-1}{m_i-1}\binom{N_i-u-m_i-1}{n_i-m_i-1} \Big/ \binom{N_i-1}{n_i-1}.$$

Therefore,

$$\text{pr}\{Y_{i,(j)} = z\} = \binom{j-1}{m_i-1}\binom{N_i-j}{n_i-m_i} \Big/ \binom{N_i-1}{n_i-1},$$

Sedransk and Meyer (1978, p. 247) find that the analogous design-based is probability is

$$\binom{j}{m_i}\binom{N_i-j}{n_i-m_i} \Big/ \binom{N_i}{n_i},$$

so we see that their confidence intervals based on the order statistics from each stratum are closely related to the Baysian prediction intervals. One advantage of the Bayes' approach is that no additional assumptions about $\text{pr}\{Z_{(t)} = Y_{i,(j)}\}$ are required.

Now, if $n_i \to \infty$, $N_i - n_i \to \infty$, then using the asymptotic normality of the Beta-Binomial distribution, (3.2) tends to

$$\text{pr}\left[W \geq \left\{\Sigma \frac{N_i(N_i-n)}{n_i} p_i(1-p_i)\right\}^{-\frac{1}{2}} (j - \Sigma N_i p_i)\right] , \qquad (3.3)$$

where W is standard normal. We note that the centre of the asymptotic distribution is at $j = \Sigma N_i p_i$, which gives some justification for using the weighted cumulative distribution function for estimating population percentiles.

Bayesian prediction intervals based on (3.3) are analogous to the large-sample confidence intervals developed by Woodruff (1952) for this case. The main difference is that in (3.3), the p_i's depend on the boundary points of the intervals, whereas Woodruff (1952) uses $p_i'(1-p_i')$ instead of $p_i(1-p_i)$, where p_i' is the proportion of observations in the i-th stratum which are less than the estimated percentile. For large samples, the two methods will give virtually identical intervals.

4. EXTENSIONS TO MORE COMPLEX DESIGNS

For estimating population means in simple or stratified random samples, the Bayesian approach which we adopted corresponds closely with the standard design-based procedures, at least asymptotically, except that the interval estimates are not confidence intervals, but are intervals with a given probability content in the posterior. These models also give tractable results for estimating population percentiles.

The extension of these models to more complex designs, such as two-stage or multi-stage sampling is not obvious. For example, if stratum i is not sampled, as in two-stage sampling, then the data give no information on the population values in that stratum, under the models we have described.

One extension of our procedures to these cases is the use of a prior which is a mixture of Dirichlet process (Antoniak, 1974; Dalal and Hall, 1980). The implications of these models deserve further study.

REFERENCES

ANTONIAK, C.E., (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist., 2,* 1152-1174.

BRETH, M., (1979), Nonparametric Bayesian interval estimation. *Biometrika, 66,* 641-644.

CAMPBELL, G. and HOLLANDER, M., (1978), Rank order estimation with the Dirichlet prior. *Ann. Statist., 6,* 142-153.

DALAL, S.R. and HALL, G.J., (1980), On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist., 8,* 664-672.

ERICSON, W.A. (1969), Subjective Bayesian models in sampling finite populations (with discussion). *J.R. Statist. Soc. B, 31,*195-233.

FERGUSON, T.S. (1973), A Bayesian analysis of some nonparametric problems. *Ann. Statist., 1,* 209-230.

———— (1974), Prior distributions on spaces of probability measures. *Ann. Statist., 2,* 615-629.

HILL, B.M. (1968), Posterior distribution of percentiles: Bayes' theorem for sampling from a finite population. *J. Amer. Statist. Assoc., 63,* 677-691.

SCOTT, A. (1971), Large sample posterior distributions for finite populations. *Ann. Math. Statist., 42,* 1113-1117.

SEDRANSK, J. and MEYER, J. (1978), Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling. *J.R. Statist. Soc. B, 40,* 239-252.

SMITH, T.M.F. (1976), The foundations of survey sampling: a review (with discussion). *J.R. Statist. Soc. A,* 139, 183-204.

WILKS, S.S. (1962), *Mathematical Statistics,* 2nd edition. New York: Wiley.

WOODRUFF, R. (1952), Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc., 47,* 635-646.

Statistics Canada Statistique Canada
Ottawa

-7
Z-113F
C.2.

MODELES BAYESIENS NON PARAMETRIQUES POUR DES ECHANTILLONS DE POPULATIONS FINIES

par

DAVID A. BINDER[1]

Décembre 1981

[1]Une partie des travaux de recherche pour cette étude ont été effectués pendant que l'auteur était à la faculté d'administration de l'Université d'Ottawa.

Résumé

A l'aide des distributions a priori non paramétriques proposées par Ferguson (1973) dans une analyse bayesienne des populations finies, on démontre que, du moins asymptotiquement, des modèles bayésiens justifient les estimations et les intervalles de confiance habituels calculés pour la moyenne de la population à partir d'échantillons aléatoires simples et stratifiés. Ces modèles sont ensuite appliqués à l'estimation de percentiles de la population et on propose une nouvelle méthode d'estimation par intervalle pour les échantillons stratifiés.

Mots clés : distributions a posteriori asymptotiques, processus de Dirichlet, moyenne de la population, percentiles de la population, échantillons stratifiés, modèles de superpopulation.

1. INTRODUCTION

Il existe actuellement un certain nombre de méthodes générales d'induction statistique à partir de données recueillies dans une population finie. Un bon nombre de ces techniques sont résumées par Smith (1976). La méthode la plus répandue en pratique est probablement la technique d'induction basée sur le plan de sondage d'échantillons aléatoires, où on utilise des critères, comme le biais et l'erreur quadratique moyenne, qui sont calculés à partir de la distribution d'échantillonnage des données. Cette méthode n'exclut pas la possibilité d'utiliser l'estimateur pour un modèle de superpopulation théorique , mais dans les intervalles de confiance et l'erreur quadratique moyenne, on tient compte seulement de la distribution

de probabilité des données en vertu du plan de sondage, abstraction faite du modèle. Une des raisons probables pour lesquelles cette technique s'est tellement répandue tient à son caractère non paramétrique et au fait que, la taille des échantillons ayant tendance à être grande, elle permet beaucoup d'applications. La perte d'efficacité inhérente à la méthode basée sur le plan d'échantillonnage, comparativement aux modèles paramétriques de superpopulation, paraît donc avoir moins d'importance que la robustesse que cette méthode représente pour un grand nombre de modèles.

Un modèle bayésien d'induction pour une population finie avec des distributions a priori échangeables a été présenté par Ericson (1969). Afin de faciliter la description de ces distributions, Ericson (1969) propose l'utilisation de modèles paramétriques de superpopulation. C'est-à-dire que, pour un paramètre vectoriel θ, on peut supposer que la population finie est la réalisation d'observations indépendantes qui ont une distribution identique définie par la fonction $(P(./\theta)$. Pour toute distribution conditionnelle a priori qui correspond à la valeur de θ, on obtient une distribution a priori échangeable pour les valeurs dans la population. Ericson (1969) noté qu'un des problèmes que soulève cette méthode est qu'elle repose sur l'hypothèse selon laquelle on possède de bons renseignements a priori sur la forme de la distribution de la population finie. Pour surmonter cette difficulté, Ericson propose que la distribution de la superpopulation soit supposée discrète, et il utilise une distribution de Dirichlet comme approximation commode de la probabilité a priori. Scott (1971) a démontré que, dans des conditions de régularité appropriées, la distribution asymptotique a posteriori des valeurs de la population finie est indépendante de la distribution a priori et qu'elle obéit à une loi normale multidimensionnelle.

Le désavantage que présente cette méthode est que les valeurs discrètes de la distribution de la superpopulation sont supposées faire partie d'un sous-ensemble d'un ensemble

dénombrable connu. Dans la présente étude, on modifie cette contrainte de façon à permettre à ces valeurs discrètes de prendre n'importe quelle valeur réelle à l'aide des distributions a priori non paramétriques présentées par Ferguson (1973), c'est-à-dire par le processus de Dirichlet. Nous examinons l'estimation de moyennes et de percentiles de la population à partir d'échantillons aléatoires simples et stratifiés.

La section 2 décrit brièvement le processus de Dirichlet et présente les estimateurs qui servent à calculer les moyennes de la population.

Dans la section 3, on analyse l'importance de ces modèles dans l'estimation de percentiles de la population. On examine également des caractéristiques analogues du modèle étudié ici et de la méthode basée sur le plan de sondage dans les études de Woodruff (1952) et de Sedransk et Meyer (1978).

On n'examinera pas ici le problème des méthodes de sélection des échantillons. Etant donné qu'on applique des principes bayesiens, il n'est pas nécessaire de connaître la distribution de probabilité du plan de sondage. Notre analyse se limite donc aux conclusions qui peuvent être tirées au sujet de la population finie d'un échantillon donné.

2. LE MODÈLE ET L'ESTIMATION DES MOYENNES DE LA POPULATION

2.1 Processus de Dirichlet

Une des descriptions les plus simples du processus de Dirichlet a été présentée par Breth (1979) :

"Soit X l'espace des valeurs possibles et $(A_1, \ldots, A_{m+1})$ un ensemble d'événements dont un et seulement un doit se produire, c'est-à-dire une partition. Écrivons $P(A_i)$ pour la probabilité aléatoire, $pr(A_i)$. On obtient un processus de Dirichlet si, pour toute partition possible de toutes les valeurs possibles de m, les probabilités aléatoires multinomiales $\{P(A_1), \ldots, P(A_m)\}$ ont une distribution de Dirichlet avec le paramètre $\{\alpha(A_1), \ldots, \alpha(A_m); \alpha(A_m); \alpha(A_{m+1})$ (Wilks, 1962, p. 178)". (Traduction libre)

Pour représenter ce processus, on utilise la notation $P \in D(\alpha)$ pour une distribution de probabilité aléatoire qui résulte d'un processus de Dirichlet. Ce processus a été étudié par Ferguson (1973, 1974). Or, si $P \in D(\alpha)$ et on observe $X_1, \ldots, X_n$, des variables aléatoires indépendantes distribuées identiquement selon la loi P, la distribution a posteriori de P est un processus de Dirichlet avec le paramètre $\alpha + \Sigma \delta_{x_i}$, où δ_x représente la mesure qui donne une densité de 1 au point x.

L'inconvénient de la distribution a priori qui suit un processus de Dirichlet est que si $P \in D(\alpha)$, il s'ensuit que, avec une probabilité de 1, P est une distribution discrète. Toutefois, dans le cas des populations finies, il ne s'agit pas d'une contrainte importante puisque, comme le souligne Ericson (1969, p. 206), presque toutes les observations sont intrinsèquement discrètes à cause des imperfections des instruments de mesure, etc. Scott (1971) considère que ce caractère discret correspond aux "réalités de la pratique".

2.2 Population simple

Supposons qu'il y a une distribution de superpopulation unique à partir de laquelle toutes les valeurs de la population sont déterminées. Soit $Y_1, \ldots, Y_N$, les valeurs de la population

pour une population finie de taille N et, pour une distribution de superpopulation donnée, P, ces valeurs sont indépendantes et distribuées identiquement suivant la loi P.

$y_1, \ldots, y_n$ sont les valeurs de l'échantillon tiré de cette population et nous examinons ici le cas où la distribution a priori de P est un processus de Dirichlet simple avec le paramètre α. Par induction, on veut obtenir des résultats au sujet de $\mu = \frac{1}{N}\sum Y_i$, étant donné les observations $y_1, \ldots, y_n$. Soit $\bar{y} = n^{-1}\sum y_i$ et $s^2 = n^{-1}\sum (y_i - \bar{y})^2$.

Abstraction faite des observations, la moyenne a priori de μ est représentée par $\hat{\mu}_o$ et la variance a priori de μ est $\sigma_o^2(\mu)$. Posons que $A = \alpha(-\infty, \infty)$, la valeur du paramètre du processus de Dirichlet pour l'ensemble de la droite réelle. On obtient comme moyenne a posteriori de μ :

(2.1)

La variance a posteriori de μ est :

(2.2)

On voit que, pour les échantillons dont la taille est grande :

(2.3)

(2.4)

Les formules (2.1) à (2.4) sont données par Ericson, (1969, p. 212, 213), qui suppose que la superpopulation est formée d'unités discrètes dont la densité de probabilité est connue.

Examinons la distribution de probabilité asymptotique a posteriori pour μ. Supposons qu'il existe k valeurs distinctes parmi les $y_1, \ldots, y_n$ et formons la partition $(A_1, \ldots, A_k)$ de manière que chaque valeur distincte corresponde à exactement un

élément des A_i. Soit z_i, la valeur de l'échantillon dans A_i, et $n(A_i) = np_i$, le nombre de valeurs de l'échantillon égales à z_i. Si Y_j^* est une valeur non échantillonnée de la population, on obtient le résultat suivant pour la distribution de probabilité a posteriori : $pr(Y_j^* = z_i \mid Y_j^* \in A_i) = n(A_i)/\{\alpha(A_i) + n(A_i)\}$. Donc, si la distribution de superpopulation est discrète, cette probabilité converge vers 1 lorsque n est grand. Pour $N(A_i)$, le nombre de valeurs de la population dans A_i, on obtient donc asymptotiquement :

Scott (1971) a démontré que, lorsque $n \to \infty$ et que $N-n \to \infty$, et pour $U_i = n^{\frac{1}{2}}N^{-\frac{1}{2}}(N-n)^{-\frac{1}{2}}\{N(A_i) - Np_i\}$, $(U_1, \ldots, U_k)$ est asymptotiquement singulier et obéit à une loi normale multidimensionnelle avec une moyenne de zéro et une matrice de covariance $\Sigma = (\sigma_{ij})$ où

Donc, $\left\{\frac{nN}{N-n}\right\}^{\frac{1}{2}}(\mu - \hat{\mu}n)$ suit une loi $N(0,s^2)$ lorsque $n \to \infty$ et que $N-n \to \infty$. Dans ce cas, les intervalles de prévision bayésiens sont asymptotiquement équivalents aux intervalles de confiance habituels.

2.3 Population stratifiée

Supposons que les valeurs de la population sont stratifiées en L strates de taille $N_1, \ldots, N_L$. Il est possible que ces strates faisaient partie du plan de sondage initial ou que les observations ont été stratifiées a posteriori. Les valeurs de la population dans la $i^{ième}$ strate sont $Y_{i1}, \ldots, Y_{iN_i}$. Les valeurs de l'échantillon sont $y_{i1}, \ldots, y_{in_i}$.

Il est supposé que, dans la $i^{ième}$ strate, les valeurs de la population sont indépendantes et distribuées identiquement selon la fonction de probabilité P_i. On examine ici le cas où la distribution a priori de $(P_1, \ldots, P_L)$ est définie de façon que les fonctions de probabilité sont indépendantes et que $P_i \in D(\alpha_i)$.

La distribution a posteriori de $(P_1, \ldots, P_L)$, étant donné $\{y_{ij}\}$, est caractérisée par le fait que les fonctions de distribution sont indépendantes et que $P_i \in D(\alpha_i + \sum_j \delta_{y_{ij}})$.

On définit $\mu_i = N_i^{-1} \sum_j Y_{ij}$ et on veut obtenir par induction des résultats sur $\mu = N^{-1} \sum N_i \mu_i$. De façon analogue à la méthode suivie à la section 2.2, on définit $A_i = \alpha_i(-\infty, \infty)$, $\bar{y}_i = n_i^{-1} \sum_j y_{ij}$ et $s_i^2 = n_i^{-1} \sum_j (y_{ij} - \bar{y}_i)^2$. L'espérance mathématique a priori de μ_i est représentée par $\hat{\mu}_{io}$ et la variance a priori de μ_i est $\sigma^2_o(\mu_i)$. L'espérance et la variance a posteriori de μ sont

et

Pour des grands échantillons, on obtient :

comme valeurs a posteriori.

En outre, lorsque $n_i \to \infty$ et que $N_i - n_i \to \infty$ on observe que

est asymptotiquement N(0,1). Là encore, pour des grands échantillons, les intervalles de prévision bayésiens correspondent aux intervalles de confiance habituels pour les échantillons stratifiés.

3. PERCENTILES DE LA POPULATION

3.1 Population simple

La présente section porte sur le problème de l'estimation des percentiles de la population.

On examine d'abord le cas où $P \in \mathcal{D}(\alpha)$ et, étant donné P, les valeurs $Y_1, \ldots, Y_N$ sont indépendantes et distribuées identiquement suivant la loi P. On ordonne les valeurs de la population et on les redésigne $Y_{(1)}, \ldots, Y_{(N)}$, de manière que $Y_{(1)} \leq \ldots \leq Y_{(N)}$. On étudie le problème de l'estimation de $Y_{(i)}$ à partir de l'échantillon $y_{(1)}, \ldots, y_{(n)}$, où $y_{(1)} \leq \ldots \leq y_{(n)}$. Campbell et Hollander (1978) ont travaillé sur un problème semblable où les distributions a priori suivent un processus de Dirichlet. Ces auteurs se sont penchés sur l'estimation du rang de $y_{(j)}$ parmi $Y_{(1)}, \ldots, Y_{(N)}$. Hill (1968) a traité ce problème au moyen d'une distribution a priori différente, et Ericson (1969) a utilisé la distribution a priori de Dirichlet pour le cas où les valeurs discrètes de la population sont connues. Les résultats présentés ici sont analogues à ceux obtenus par Ericson (1969) pour la distribution a posteriori exacte. Ils sont généralisés au cas asymptotique. Sedrask et Meyer (1978) ont étudié ce problème de manière non bayesienne, en calculant des intervalles de confiance de forme $y_{(j)} \leq Y_{(i)} \leq y_{(k)}$. Woodruff (1952) a élaboré des intervalles de confiance de $Y_{(i)}$ pour des grands échantillons.

Or, la distribution a priori de $Y_{(i)}$ dans notre modèle basé sur le processus de Dirichlet est la suivante :

où U(y) est le nombre de valeurs dans la population qui sont égales ou inférieures à y. Si on définit $A(y) = \alpha(-\infty, y]$ et $A = \alpha(-\infty, \infty)$, U(y) a alors une distribution binomiale bêta $[N; A(y), A-A(y)]$.

La distribution binomiale bêta $(M; \beta 1, \beta 2)$ est exprimée par la formule :

Supposons qu'on observe $y_{(1)}, \ldots, y_{(n)}$. Par principe, on définit $y_{(o)} = -\infty$ $Y^*_{(1)}, \ldots, Y^*_{(N-n)}$ sont les statistiques d'ordre des unités non échantillonnées. Pour $j<i$ et $y_{(j)} \leq y < y_{(j+1)}$,

(3.1)

où U(y) suit une loi binomiale bêta $\{N-n;\ A(y) + j,\ A-A(y) + n-j\}$. Cette expression donne la distribution a posteriori exacte de $Y_{(i)}$. Ericson (1969) a démontré que, dans le cas de lois a priori diffuses, c'est-à-dire quand A(y) = 0, les intervalles de prévision bayesiens correspondent de près aux intervalles de confiance classiques, à condition que N soit grand en comparaison de n.

Si $n \to \infty$, $N-n \to \infty$ et $j/n \to p$, il découle de la normalité asymptotique de la distribution binomiale bêta, qu'on obtient dans un cas particulier du théorème présenté par Scott (1971), que l'expression (3.1) devient asymptotiquement ainsi :

, où W est une variable normale réduite.

Les intervalles de prévision produits par cette méthode sont semblables aux intervalles de confiance obtenus par Woodruff (1952) pour les grands échantillons. Ces deux types d'intervalles seraient identiques si p(1-p) dans le premier terme était remplacé par $i(N-i)/N^2$. Dans le cas des grands échantillons, la différence entre ces deux genres d'intervalles serait négligeable.

3.2 Populations stratifiées

Sedransk et Meyer (1978) ont examiné le problème du calcul d'intervalles de confiance pour les percentiles de la population à partir d'échantillons aléatoires stratifiés. Ces auteurs analysent en détail le cas où L = 2 et ils proposent un certain nombre de méthodes dont la qualité des résultats est évaluée au moyen d'une étude de Monte Carlo. Cette technique devient très complexe lorsqu'il y a plus de deux strates.

Comme à la section 2.3, il est supposé que $P_i \in \mathcal{D}(\alpha_i)$ et que les $(P_1, \ldots, P_L)$ sont indépendants. On obtient que les variables $\{Y_{ij};\ 1 \leq i \leq L,\ 1 \leq j \leq N_i\}$, étant donné les valeurs de $(P_1, \ldots, P_L)$, sont indépendantes et la distribution de Y_{ij} est P_i. Dans la $i^{\text{ième}}$ strate, les observations sont $y_{i1}, \ldots, y_{in_i}$.

On définit $Z_{(1)}, Z_{(2)}, \ldots, Z_{(N_1 + \ldots + N_L)}$ comme les statistiques d'ordre de toutes les valeurs de la population, abstraction faite de la stratification, $Z_{(1)} \leq Z_{(2)} \leq \ldots Z_{(N_1 + \ldots + N_L)}$. On définit $y_{i,(1)}, \ldots, y_{i,(n_i)}$ comme les statistiques d'ordre des valeurs de la $i^{\text{ième}}$ strate, et $Z^*_{(1)}, \ldots, Z^*_{(\sum N_i - \sum n_i)}$ comme celles des unités non échantillonnées. L'objectif est de trouver la probabilité a posteriori que $Z_{(j)} \leq z$, étant donné toutes les valeurs de y. On définit $U_i(z)$ comme le nombre des valeurs de la population dans la strate i, qui sont inférieures ou égales à z. Donc :

Pour une valeur donnée de z, on définit $p_1, \ldots, p_L$ de façon que

Autrement dit, p_i est la proportion d'observations dans la $i^{ième}$ strate qui sont inférieures ou égales à z. Dans la distribution a posteriori, pour $j \geq n_i p_i$:

(3.2)

où $U_1, \ldots, U_L$ sont indépendants et

U_i suit une loi binomiale bêta $\{$ $\}$.

Comparons maintenant les intervalles de prévision pour petits échantillons et les intervalles de confiance des percentiles de la population élaborés par Sedransk et Meyer (1978). On examinera en particulier le cas où il y a L-2 strates et, pour simplifier l'analyse, on supposera, comme Ericson (1969), que toutes les valeurs de l'échantillon sont distinctes et que $A(y) \doteq 0$. On définit $Y_{i,(j)}$ comme la $j^{ième}$ statistique d'ordre des observations de la $i^{ième}$ strate dans l'expression $pr\{Z_{(t)} \leq z \mid Z_{(t)} = Y_{i,(j)}\}$. Une fois ces valeurs calculées, les intervalles de confiance sont déterminées en fonction des diverses hypothèses adoptées au sujet des valeurs des strates dans le calcul de $pr\{Z_{(t)} = Y_{i,(j)}\}$.

Ensuite, on calcule la probabilité a posteriori de Bayes pour l'événement $Y_{i,(j)} \leq z$, où $z = y_{i,(m_i)}$. En suivant les étapes décrites plus haut, on voit que cette probabilité est identique à $pr\{U_i(z) \geq j - m_i\}$, où U_i suit une distribution binomiale bêta $\{N_i - n_i;\ m_i,\ n_i - m_i\}$. Donc :

et

D'après Sedransk et Meyer (1978, p. 247), la probabilité correspondante évaluée par la méthode basée sur le plan de sondage est

ce qui nous permet de voir que les intervalles de confiance basés sur les statistiques d'ordre à l'intérieur de chaque strate correspondent à peu près aux intervalles de prévision bayésiens. Un des avantages de la méthode bayésienne est qu'il n'est pas nécessaire d'adopter d'autres hypothèses concernant $\text{pr}\{Z_{(t)} = Y_{i,(j)}\}$.

Si $n_i \to \infty$ et $N_i - n_i \to \infty$, le membre de droite de l'équation (3.2) se transforme, en vertu de la normalité asymptotique de la distribution binomiale bêta, en l'expression suivante :

(3.3)

où W obéit à une loi normale réduite. On note que le centre de la distribution asymptotique se trouve à $j = \Sigma N_i p_i$, ce qui justifie en partie l'utilisation de la forme pondérée de la fonction cumulative de distribution dans l'estimation des percentiles de la population.

Les intervalles de prévision bayesiens basés sur l'expression (3.3) sont analogues aux intervalles de confiance pour grands échantillons calculés par Woodruff (1952), dans le cas de populations stratifiées. La principale différence est que, dans l'expression (3.3), les p_i varient en fonction des points limites des intervalles, alors que Woodruff (1952) utilise $p_i'(1-p_i')$ au lieu de $p_i(1-p_i)$, p_i' étant la proportion d'observations, dans la $i^{\text{ième}}$ strate, qui sont inférieures au percentile estimé. Dans le cas de grands échantillons, ces deux méthodes fournissent des intervalles pratiquement identiques.

4. GENERALISATION AUX PLANS DE SONDAGE COMPLEXES

Pour l'estimation de moyennes de la population à partir d'échantillons aléatoires simples ou stratifiés, la méthode bayesienne adoptée dans cette étude correspond de près aux techniques habituelles basées sur le plan de sondage, du moins asymptotiquement, sauf que les intervalles estimés sont non pas des intervalles de confiance, mais des intervalles auxquels une probabilité est attribuée par la distribution a posteriori. Ce genre de modèle produit aussi des résultats utiles pour l'estimation de percentiles de la population.

L'application de ces modèles à des plans de sondage plus complexes, comme dans le cas d'échantillons à deux ou à plusieurs degrés, n'est pas directe. Par exemple, si la strate i n'est pas échantillonnée, comme dans les échantillons à deux degrés, les données ne fournissent aucun renseignement sur les valeurs de la population dans cette strate, si on applique les modèles décrits dans la présente étude.

Une façon d'appliquer nos modèles aux plans de sondage complexes serait d'inclure une distribution a priori mixte qui renferme un processus de Dirichlet (Antoniak, 1974, Dalal et Hall, 1980). L'utilité de ce genre de modèle mérite d'être analysée plus en détail.

BIBLIOGRAPHIE

ANTONIAK, C.E., (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist., 2, 1152-1174.

BRETH, M., (1979), Nonparametric Bayesian interval estimation. Biometrika, 66, 641-644.

CAMPBELL, G. et HOLLANDER, M., (1978), Rank order estimation with the Dirichlet prior. Ann. Statist., 6, 142-153.

DALAL, S.R. et HALL, G.J., (1980), On approximating parametric Bayes models by nonparametric Bayes models. Ann. Statist., 8, 664-672.

ERICSON, W.A. (1969), Subjective Bayesian models in sampling finite populations (avec analyse). J.R. Statist. Soc. B, 31, 195-233.

FERGUSON, T.S. (1973), A Bayesian analysis of some nonparametric problems. Ann. Statist., 1, 209-230.

_____ (1974), Prior distributions on spaces of probability measures. Ann. Statist., 2, 615-629.

HILL, B.M. (1968), Posterior distribution of percentiles : Bayes' theorem for sampling from a finite population. J. Amer. Statist. Assoc., 63, 677-691.

SCOTT, A. (1971), Large sample posterior distributions for finite populations. Ann. Math. Statist., 42, 1113-1117.

SEDRANSK, J. et MEYER, J. (1978), Confidence intervals for the quantiles of a finite population : simple random and stratified simple random sampling. J.R. Statist. Soc. B, 40, 239-252.

SMITH, T.M.F. (1976), The foundations of survey sampling : a review (avec analyse). J.R. Statist. Soc. A, 139, 183-204.

WILKS, S.S. (1962), Mathematical Statistics, 2nd edition. New York : Wiley.

WOODRUFF, R. (1952), Confidence intervals for medians and other position measures. J. Amer. Statist. Assoc., 47, 635-646.

POOR

QUALITY

ORIGINAL